

AD 684087

AD

U S NAVAL
PERSONNEL RESEARCH ACTIVITY

ANALYSIS OF PERSONNEL DATA

RESEARCH MEMORANDUM SRM 69-17

MARCH 1969

PATTERN CLUSTERING BY MULTIVARIATE MIXTURE ANALYSIS

John H. Wolfe

THIS DOCUMENT HAS BEEN APPROVED FOR PUBLIC
RELEASE AND SALE; ITS DISTRIBUTION IS UNLIMITED

D D C

RECEIVED
MAR 21 1969
RECEIVED

Reproduced by the
CLEARINGHOUSE
for Federal Scientific & Technical
Information Springfield Va. 22151



AN ACTIVITY OF THE BUREAU OF NAVAL PERSONNEL

**Best
Available
Copy**

AD

PATTERN CLUSTERING BY MULTIVARIATE MIXTURE ANALYSIS

By

John H. Wolfe

IER160000A03

Research Report SRM 69-17

Submitted by

Richard C. Sorenson, Ph.D., Director, Statistical Department

Approved by

E.E. Dudek, Ph.D., Technical Director

G.W. Watson, Commander, USN

Commanding Officer

This document has been approved for public release
and sale; its distribution is unlimited.

U. S. Naval Personnel Research Activity
San Diego, California 92152

SUMMARY

A. Problem

This report is addressed to the problem of developing procedures for clustering individuals or objects into similar "types." Such procedures could be useful in producing an objective analysis and revision of the Navy rating structure by classifying positions with similar patterns of skill requirements into the same occupational category.

B. Background and Requirements

1. Cluster analysis methods of various kinds have been employed in the study of individual differences, in the taxonomy of biological organisms, in the classification of documents for information retrieval, in the study of Navy enlisted basic skill patterns, and in unsupervised pattern recognition of electronic signal patterns.

2. Recent interest in this area has been stimulated by the advent of high-speed digital computers capable of carrying out cluster analysis automatically. However, most existing methods contain certain arbitrary factors or assumptions which are difficult to justify statistically. Since different methods can give different results, there is a clear requirement for the development of a technique which is rigorously derived from statistical theory. Such a technique was presented by Wolfe (1965, 1967) for the special case of mixtures of multivariate normal distributions. The present report generalizes these methods to other distributions.

C. Approach

The approach involves reformulating cluster analysis as a problem in the estimation of the parameters of a mixture of distributions. Maximum-likelihood (ML) methods are used exclusively because of the ease with which they can be generalized to multivariate distributions of various forms.

D. Findings

1. Regardless of the shape of the distribution the maximum-likelihood estimate of the proportion of a mixture from a given type is equal to the sample mean of the probability of membership of the objects in that type. The equations for the maximum-likelihood estimates of the parameters of a mixture are the weighted averages of the expressions used in obtaining ML estimates for pure types, where the weights are the probabilities of membership.

2. The estimation procedures for normal mixtures with unequal covariances, normal mixtures with equal covariances, and mixtures of latent classes are derived as special cases of the general theory.

3. Various iteration techniques are discussed for obtaining numerical solutions to mixture problems.

4. The examples of the results of two computer mixture analysis programs (NORMIX and NORMAP) indicate that the theory is sound for large samples and that the procedures given in this paper are practical.

E. Conclusion

A practical and statistically rigorous method of cluster analysis has been developed.

F. Recommendations

1. It is recommended that the computer programs NORMIX and NORMAP be used in Naval research studies requiring a cluster analysis of continuous measurement patterns.

2. Further development of computer programs for clustering discrete data patterns is desirable.

REPORT USE AND EVALUATION

Feedback from consumers is a vital element in improving products so that they better respond to specific needs. To assist the Chief of Naval Personnel in future planning, it is requested that the use and evaluation form on the reverse of this page be completed and returned. The page is preaddressed and franked; fold in thirds, seal with tape, and mail.

DEPARTMENT OF THE NAVY

Postage and Fees Paid
Navy Department

Official Business

Chief of Naval Personnel (Pers-A3)
Department of the Navy
Washington, D.C. 20370

Report Title & No: PATTERN CLUSTERING BY MULTIVARIATE MIXTURE ANALYSIS
(SRM69-17)

1. Evaluation of Report. Please check appropriate column.

FACTORS	RATING			COMMENTS
	LOW	AVE	HIGH	
Usefulness of Data				
Timeliness				
Completeness				
Technical Accuracy				
Validity of Recommendations				
Soundness of Approach				
Presentation and Style				
Other				

2. Use of Report. Please fill in answers as appropriate.

- What are your main uses for the material contained in the report?
- What changes would you recommend in report format to make it more useful?
- What types of research would be most useful to you for the Chief of Naval Personnel to conduct?
- Do you wish to remain on our distribution list?
- Please make any general comments you feel would be helpful to us in planning our research program.

NAME: _____ CODE: _____

ORGANIZATION: _____

ADDRESS: _____

CONTENTS

	Page
Summary & Conclusions.	iii
1. INTRODUCTION	1
2. GENERAL MIXTURE ANALYSIS	4
3. ITERATION METHODS.	8
4. MULTIPLE SOLUTIONS AND INITIAL ESTIMATES	12
5. NORMAL MIXTURE ANALYSIS.	14
6. LATENT CLASS ANALYSIS.	16
7. EXAMPLES OF COMPUTER ANALYSES.	17
7.1 Iris Data	18
7.2 Artificial Clusters	22
8. CONCLUDING REMARKS	25

TABLES

1. Fisher-Iris Significance Tests	19
2. Fisher-Iris Parameters	21
3. Artificial Cluster Significance Tests.	23
4. Artificial Cluster Parameters.	24
REFERENCES	27

PATTERN CLUSTERING BY MULTIVARIATE MIXTURE ANALYSIS

1. INTRODUCTION

This paper is addressed to the problem which has been variously called cluster analysis, Q-analysis, typology, grouping, clumping, classification, numerical taxonomy, and unsupervised pattern recognition. The variety of nomenclature may be due to the importance of the subject in such diverse fields as psychology, biology, signal detection, artificial intelligence, and information retrieval. Perhaps this multiplicity of names also indicates a certain confusion in the basic definition of the problem. This paper attempts to clarify the formulation of the problem, with a resulting improvement in conceptual simplicity and statistical rigor.

In classification methodology, one is generally given a sample of N objects or individuals, each of which is measured on m variables. From this information alone, one must devise a classification scheme for grouping the objects into r classes. The number of classes and the characteristics of the classes are to be determined. If all the objects in a given class were identical to one another, the problem would be simple. However, in the usual situation the objects in a class differ on most or all of the measures. Most cluster analysis procedures try to measure the "similarity" of objects within a class, and then try to group the objects so as to maximize within-class similarity. Unfortunately, the appropriate measure of similarity is a subject of some controversy. It

would be desirable to derive a cluster analysis system without arbitrary assumptions about similarity. Such a system will be presented in this paper.

Since the objects within a class differ from one another, it is reasonable to assume the existence of a probability distribution of characteristics for a population belonging to this class. Elements of a different class will have a different probability distribution of characteristics. The combined population taken from all classes will have a probability distribution which is a mixture of distributions. The problem is to identify and describe the component distributions from a sample drawn from the mixture. Before it is possible to solve this problem, some assumptions must be made about the forms of the component distributions. For example, the component distributions are usually assumed to be unimodal. The purpose of classification methodology is to take a complicated multi-modal distribution and analyze it into simple familiar components. Therefore, the component distributions can usually be assumed to be standard statistical distributions with unknown parameters. The classification problem can then be solved by standard statistical techniques of parametric estimation. This is the approach taken in the present paper.

Over 70 years ago Karl Pearson (1894) used the method of moments to estimate the parameters of a mixture of two univariate normal distributions. Maximum-likelihood methods for a special case of the same problem were presented by Rao (1952). Studies of mixtures of univariate discrete distributions have been reviewed by Blischke (1963). Maximum-likelihood (ML)

estimation procedures for mixtures of multivariate normal distributions were presented by Wolfe (1965,1967). Similar ML estimation methods were presented by Hasselblad (1966) for univariate normals, and Cohen (1967) developed simplified moment estimators for the univariate normal case. Moment estimators for various special cases of mixtures of multivariate normals have been presented by Cooper (1967). Stanat (1968) and Sammon (1968) developed multivariate generalizations of Medgyessy's (1961) methods for estimating the parameters of a mixture from a Fourier approximation to the sample distribution.

Lazarsfeld's "Latent Structure Analysis" (1959) is closely related to the mixture analysis problem. In "Latent Class Analysis," the observed contingencies among several dichotomous variables are explained by assuming the population is a mixture of "latent classes" within each of which the variables are independently distributed. Gibson (1959) succeeded in generalizing Lazarsfeld's model to mixtures of spherical multivariate normal distributions.

This paper summarizes this author's previous work on mixtures of multivariate normal distributions and generalizes the theory to mixtures of multivariate distributions of almost any given form. The approach is exclusively that of maximum-likelihood estimation. Although other procedures are valuable in special cases, maximum-likelihood methods seem to be the

easiest to generalize and the most efficient in the way they use the information in the sample. They probably have been overlooked in the past because of the amount of computation required to solve the ML equations numerically. Computation costs are no longer prohibitive with modern electronic computers, as we shall illustrate by presenting the ML solution obtained in one minute to the classic Fisher Iris mixture problem.

2. GENERAL MIXTURE ANALYSIS

Let $\alpha_1(x, \theta_1), \alpha_2(x, \theta_2), \dots, \alpha_r(x, \theta_r)$ be r probability distributions defined on an m -dimensional space of random vectors:

$$x = (X_1, X_2, \dots, X_m) \quad .$$

Assume each α_s is a twice differentiable function of its parameters,

$$\theta_s = (\theta_{s1}, \theta_{s2}, \dots, \theta_{sq}) \quad .$$

Suppose a mixture of distributions is formed by taking proportions $\{\lambda_s\}$ of the population from types $\{\alpha_s\}$. The probability distribution of the mixture is given by

$$f(x) = \sum_{s=1}^r \lambda_s \alpha_s(x, \theta_s), \quad (2.1)$$

where

$$\sum_{s=1}^r \lambda_s = 1 \quad . \quad (2.2)$$

The "probability of membership" of a vector x in type s can be defined as

$$P(s|x) = \frac{P(s) \cdot P(x|s)}{P(x)} = \frac{\lambda_s \alpha_s(x, \theta_s)}{f(x)} \quad . \quad (2.3)$$

Suppose a sample of N random vectors is drawn from the mixture.
The k^{th} random vector is represented by

$$x_k = (x_{1k}, x_{2k}, \dots, x_{mk})$$

The maximum-likelihood estimates of the parameters are those values of $\{\lambda_s, \theta_s\}$ which maximize the likelihood of the sample,

$$\log L = \sum_{k=1}^N \log f(x_k),$$

subject to the constraint $\sum_{s=1}^r \lambda_s = 1$.

Using a Lagrangian multiplier ω , we form the function

$$\log L' = \sum_{k=1}^N \log f(x_k) - \omega \left(\sum_{s=1}^r \lambda_s - 1 \right) \quad (2.4)$$

The necessary equations for maximum likelihood are obtained by setting the derivatives of $\log L'$ to zero as follows:

$$\frac{\partial \log L'}{\partial \lambda_s} = \sum_{k=1}^N \frac{1}{f(x_k)} \alpha_s(x_k) - \omega = 0, \quad (2.5)$$

$$\text{and } \frac{\partial \log L'}{\partial \theta_{si}} = \sum_{k=1}^N \frac{\lambda_s}{f(x_k)} \frac{\partial(\alpha_s)}{\partial \theta_{si}} = 0 \quad (2.6)$$

Multiplying (2.5) by λ_s , and substituting from (2.3), we have

$$\lambda_s \frac{\partial \log L'}{\partial \lambda_s} = \sum_{k=1}^N P(s|x_k) - \omega \lambda_s = 0 \quad (2.7)$$

Summing across s and using (2.1) and (2.2), we find that $\omega=N$.

After a little algebra on equation (2.7), we obtain the following:

Theorem 1 The maximum-likelihood estimate of the proportion of a mixture from a given type is equal to the sample mean of the probability of membership of the objects in that type; and the likelihood equation is given by:

$$\hat{\lambda}_s = \frac{1}{N} \sum_{k=1}^N P(s|x_k) \quad (2.8)$$

The concept of probability of membership also helps to clarify equation (2.6). When substitution is made in (2.6) from (2.3), the result is the following:

Theorem 2 The equations for the maximum-likelihood estimates of the parameters of the distributions comprising a mixture are given by:

$$\frac{\partial \log L}{\partial \theta_{si}} = \sum_{k=1}^N P(s|x_k) \cdot \frac{\partial \log \alpha_s}{\partial \theta_{si}} = 0. \quad (2.9)$$

If the entire population were drawn from one type, α_s , the equations for the maximum-likelihood estimates of θ_s would be

$$\sum_{k=1}^N \frac{\partial \log \alpha_s}{\partial \theta_{si}} = 0. \quad \text{Thus the equations for the ML}$$

estimates of the parameters of a mixture are the weighted averages of the expressions used in obtaining ML estimates for pure types, where the weights are the probabilities of membership.

Usually the number of types, r , is only a hypothesis which can be tested against an alternative hypothesis of r' types by finding maximum-

likelihood estimates under both hypotheses and testing the likelihood ratio by the formula $\chi^2 = -2 \log (L_r/L_{r'})$ with degrees of freedom equal to the difference in the number of parameters estimated. Likelihood ratio tests may also be used to test alternative hypotheses concerning the forms of the component distributions. Of course the distribution of the logarithm of the likelihood ratio is approximately chi-square for large samples only.

In most cases, equations (2.8) and (2.9) will have to be solved numerically. For this purpose and also for the purpose of obtaining confidence intervals, it is desirable to have some approximation to the information matrix:

$$I = \left(\begin{array}{c|c} I_{\lambda\lambda} & I_{\lambda\theta} \\ \hline I_{\theta\lambda} & I_{\theta\theta} \end{array} \right) . \quad (2.10)$$

where I is partitioned into the sub-matrices $I_{\lambda\lambda}, I_{\lambda\theta}, I_{\theta\lambda}, I_{\theta\theta}$, defined as

$$I_{\lambda\lambda} = \left\{ E \left(\frac{\partial \log L}{\partial \lambda_s} \frac{\partial \log L}{\partial \lambda_p} \right) \right\} = N \left\{ \frac{1}{\lambda_s \lambda_p} E \left(P(s|x) P(p|x) \right) \right\} \quad (2.11)$$

$$I_{\lambda\theta} = \left\{ E \left(\frac{\partial \log L}{\partial \lambda_s} \frac{\partial \log L}{\partial \theta_{pj}} \right) \right\} = N \left\{ \frac{1}{\lambda_s} E \left(P(s|x) P(p|x) \frac{\partial \log \alpha_p}{\partial \theta_{pj}} \right) \right\} \quad (2.12)$$

$$I_{\theta\theta} = \left\{ E \left(\frac{\partial \log L}{\partial \theta_{si}} \frac{\partial \log L}{\partial \theta_{pj}} \right) \right\} = N \left\{ E \left(P(s|x) P(p|x) \frac{\partial \log \alpha_s}{\partial \theta_{si}} \frac{\partial \log \alpha_p}{\partial \theta_{pj}} \right) \right\} \quad (2.13)$$

The submatrix $I_{\theta\lambda}$ is the transpose of $I_{\lambda\theta}$. Multidimensional confidence ellipsoids can be developed with the help of the inverse of the information matrix

$$V = I^{-1} . \quad (2.14)$$

which gives the large-sample dispersions of the ML estimators. The ML equations (2.8) and (2.9) can be solved iteratively by the "method of scoring" (Kale, 1962) using the following equations:

$$\begin{pmatrix} \{\Delta\lambda_s\} \\ \{\Delta\theta_{si}\} \end{pmatrix} = V \cdot \begin{pmatrix} \left\{ \frac{1}{\lambda_p} \sum_{k=1}^N P(p|x_k) - N \right\} \\ \left\{ \sum_{k=1}^N P(p|x_k) \frac{\partial \log \alpha_p}{\partial \theta_{pj}} \right\} \end{pmatrix} \quad (2.15)$$

where $\{\Delta\lambda_s\}$ and $\{\Delta\theta_{si}\}$ are column vectors for the increments in the estimates used in the next numerical iteration.

In most cases the expectations in the information matrix involve integrals which are impossible to evaluate in closed form and difficult to approximate by series. The usual approach has been to estimate the information matrix from the sample, replacing the expectation symbol E in (2.11-2.13) by $\frac{1}{N} \sum_{k=1}^N$. This approach is satisfactory if the information matrix is to be calculated only once for the purpose of obtaining confidence regions. It is prohibitively expensive if I has to be re-estimated many times during the iteration (2.15). Some alternative iteration techniques will be developed in the next section.

3. ITERATION METHODS

Fortunately certain approximations are often possible when solving mixture problems numerically. First let us consider a limiting case. When the component types of a mixture are widely separated, each point will have a probability of membership close to unity for one of the types

and nearly zero for the other types. In other words, the probabilities of membership come close to defining a partition of the sample points into discrete clusters. The product of two probabilities of membership can be then approximated by

$$P(s|x)P(p|x) \sim \delta_{ps} P(s|x), \quad (3.1)$$

where δ_{ps} is the Kronecker delta.

When this approximation is inserted into the information matrix the result is

$$\begin{aligned} I_{\lambda\lambda} &\sim N \left\{ \frac{\delta_{ps}}{\lambda_s} \right\} \\ I_{\lambda\theta} &\sim 0 \\ I_{\theta\theta} &\sim N \left\{ \delta_{ps} E \left(\frac{\partial \log \alpha_s}{\partial \theta_{si}} \frac{\partial \log \alpha_s}{\partial \theta_{sj}} P(s|x) \right) \right\}. \end{aligned}$$

The information matrix for the mixing proportions is seen to be approximately diagonal, and since $I_{\lambda\theta} \sim 0$, the iteration for the mixing proportions can be carried out independently of the iteration for the other parameters. The approximation for $I_{\theta\theta}$ is seen to be

$$I_{\theta\theta} \sim \{N \delta_{ps} I_s\},$$

where I_s is the information matrix of the parameters $\{\theta_{si}\}$ for a single observation from a pure distribution, α_s . Thus, if the distributions do not overlap very much, the iterations for the parameters of one type do not involve terms from the other types.

Equation (2.15) then reduces to the two approximations:

$$\Delta \lambda_s = \frac{1}{N} \sum_{k=1}^N P(s|x_k) - \lambda_s. \quad (3.2)$$

$$\left\{ \Delta_{si} \right\} = I_s^{-1} \cdot \left\{ \frac{1}{N} \sum_{k=1}^N P(s|x_k) \frac{\partial \log \alpha_s}{\partial \theta_{sj}} \right\}. \quad (3.3)$$

The first equation can be put in the form

$$\lambda'_s = \lambda_s + \Delta \lambda_s = \frac{1}{N} \sum_{k=1}^N P(s|x_k),$$

where λ'_s is the estimate of $\hat{\lambda}_s$ on the next iteration. This is precisely the same form as (2.8), indicating that a good iteration technique is simply to calculate the right hand side of (2.8) with old estimates of the parameters and thus obtain improved estimates on the left side. For certain distributions (3.3) also becomes extremely simple.

The approximations introduced from (3.1) do not affect the final solution of the likelihood equations, they only affect the rate of convergence (or lack of it) in the iteration process. Even when there is an appreciable amount of overlap, the information matrix often is dominated by its diagonal to a sufficient degree that the simplified iteration methods remain useful.

The preceding results have been derived for an extreme case; let us see what can be derived under a less restrictive assumption. The fact that many of the off-diagonal elements of I tend to vanish in the extreme case of non-overlapping distributions makes it plausible to assume that in many cases of interest, the diagonal will dominate the matrix, i.e.

$$|I_{ii}| > \sum_{j \neq i} |I_{ij}| \quad \text{for every } i. \quad (3.4)$$

When this assumption holds, Issacson and Keller (1966,p.122) show that it is possible to accelerate convergence of the iteration process without explicitly calculating any part of the information matrix. The most successful iteration method for mixture problems empirically tested as of this writing has been a modification of Aitkens' acceleration process applied to equations (2.8) and (2.9) as if they were independent.

Let $\lambda_s, \lambda'_s, \lambda''_s$ represent three successive iterations with equation (2.8) and

$$\text{let } D = \max_{1 \leq s \leq r} \left(1 - \frac{\lambda''_s - \lambda'_s}{\lambda'_s - \lambda_s} \right). \quad (3.5)$$

Then the accelerated estimate of $\hat{\lambda}_s$ on the next iteration is defined as

$$\lambda'''_s = \frac{1}{D} (\lambda''_s - \lambda'_s) + \lambda''_s. \quad (3.6)$$

Suppose some simplified iteration scheme (such as (3.3)) has been derived from (2.9) for the parameters $\{\hat{\theta}_{si}\}$ and let $\theta_{si}, \theta'_{si}, \theta''_{si}$ represent three successive iterations.

$$\text{Let } D_{si} = 1 - \frac{\theta''_{si} - \theta'_{si}}{\theta'_{si} - \theta_{si}}. \quad (3.7)$$

Then the accelerated estimate of $\hat{\theta}_{si}$ for the next iteration is defined as

$$\theta'''_{si} = \frac{1}{D_{si}} (\theta''_{si} - \theta'_{si}) + \theta''_{si}. \quad (3.8)$$

It is helpful to place some minimum value (such as .05) on the values that D_{s_i} and D are allowed to attain, and to prevent D from taking on a value so small that some λ_s''' becomes negative or greater than one.

The above methods all require initial estimates to lie within some radius of convergence of the true maximum-likelihood values; otherwise iterations will diverge. The methods discussed represent only a few of a large number of possible numerical techniques; it seems quite possible that better methods for mixture problems will be developed.

4. MULTIPLE SOLUTIONS AND INITIAL ESTIMATES

The likelihood equations defined by Theorems 1 and 2 have a large number of solutions corresponding to absolute maxima, relative maxima, saddle points and minima of the likelihood function. It is easy to see that there are at least $r!$ absolute maxima, because if T is any permutation of the integers $1, 2, \dots, r$ and $\{\lambda_s, \theta_s\}$ is an absolute maximum-likelihood solution, then another absolute maximum is defined by the mapping

$$\begin{aligned}\lambda_s &\rightarrow \lambda_{T(s)} , \\ \theta_s &\rightarrow \theta_{T(s)} .\end{aligned}$$

Between the absolute maxima will be relative minima and saddle points, which also satisfy the likelihood equations. For example, if an absolute maximum-likelihood solution is available for $r-1$ types, then $r-1$ families of saddle-ridge solutions for r types can be generated by setting the parameters of the r^{th} type equal to those of one of the other types. If two types have identical parameters ($\theta_r = \theta_s$), then the likelihood remains unchanged as long as $\lambda_s + \lambda_r = \text{constant}$. Additional solutions can be generated

by restricting three of the types to have identical parameters, then four of the types, then two sets of two types, and so forth. These degenerate solutions are easily spotted, and it is not difficult to program a computer to avoid them as well as the relative minimum likelihood solutions.

There are many other relative maxima which are more difficult to identify as such. If an absolute maximum-likelihood solution is available for r types, then one way of generating initial estimates for an $r-1$ type solution is to combine two of the clusters into a larger one with greater dispersions and a centroid somewhere between the two component clusters. There are $\binom{r}{2}$ ways of selecting two out of r clusters. Not all of these different initial estimates will result in different solutions after iteration, but experience has shown that some of them will. Given two or more solutions, it is always possible to choose the "best" by selecting the one with the largest likelihood. Unfortunately, there is no algorithm for generating all possible solutions or for proving that a given solution is an absolute maximum.

Most other cluster-seeking algorithms suffer from the same difficulty: they can converge on a sub-optimal solution. Ball (1967) and Friedman and Rubin (1967) report they can obtain different solutions depending on the initial estimates. The practical answer is to try a variety of initial estimates on any given problem and to use various heuristics for generating plausible initial estimates. Some of these heuristics are Ball and Hall's "cluster splitting," and "cluster lumping," Friedman and Rubin's "forcing passes," Forgy's "reassignment passes (1965)," and random partitioning. A strong case can be made for including a human being in the system for generating initial estimates. Ball and Hall's PROMENADE (1967) uses

CRT display consoles to produce on-line graphical representation of sample scatter. The console operator can rotate the display through several dimensions and can select initial estimates by pointing to trial cluster centroids with a pointing device.

The generation of initial estimates is a whole field in itself. We will not pursue it further in this paper but will merely assume that initial estimates for the iterative maximum-likelihood procedures have been supplied by some other source. All of the various clustering techniques developed by other investigators are valuable potential sources of initial estimates.

5. NORMAL MIXTURE ANALYSIS

To illustrate the general principles of ML estimation for mixtures, let us consider the case of mixtures of multivariate normal distributions.

Let
$$\alpha_s(x, \mu_s, \sigma^s) = (2\pi)^{-\frac{m}{2}} |\sigma_s|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (x - \mu_s)' \sigma_s^{-1} (x - \mu_s) \right\}, \quad (5.1)$$

where $\sigma^s = \{\sigma_{ij}^s\}$ and $\sigma_s^{-1} = \{\sigma_s^{ij}\}$.

The derivatives are given as follows:

$$\frac{\partial \log \alpha_s}{\partial \mu_{si}} = \sum_{j=1}^m \sigma_s^{ij} (x_j - \mu_{sj}), \quad (5.2)$$

and
$$\frac{\partial \log \alpha_s}{\partial \sigma_{ij}^s} = \left(1 - \frac{\delta_{ij}}{2}\right) (\sigma_{ij}^s - (x_i - \mu_{si})(x_j - \mu_{sj})). \quad (5.3)$$

Substituting the derivatives in (2.9) and doing a little algebra, we find the following:

Theorem 3 The likelihood equations for the parameters of a mixture of multivariate normal distributions with unequal covariance matrices are equivalent to the following equations:

$$\hat{\lambda}_s = \frac{1}{N} \sum_{k=1}^N \hat{P}(s|x_k) \quad (5.4)$$

$$\hat{\mu}_{si} = \frac{1}{N\hat{\lambda}_s} \sum_{k=1}^N \hat{P}(s|x_k) X_{ik} \quad (5.5)$$

$$\hat{\sigma}_{ij}^s = \frac{1}{N\hat{\lambda}_s} \sum_{k=1}^N \hat{P}(s|x_k) (X_{ik} - \hat{\mu}_{si})(X_{jk} - \hat{\mu}_{sj}) \quad (5.6)$$

Thus the equations for estimating the parameters of a mixture of normal distributions are closely analogous to the equations for estimating the parameters of a pure type except that each sample point is weighted by its probability of membership.

When the types have a common covariance matrix σ , so that $\sigma^s = \sigma$ for $1 \leq s \leq r$, then equation (2.9) is no longer valid, since it was derived under the assumption that the parameters of different types were not functionally related. The correct likelihood equation is

$$\frac{\partial L}{\partial \sigma_{ij}} = \sum_{s=1}^r \frac{\partial L}{\partial \sigma_{ij}^s} \frac{\partial \sigma_{ij}^s}{\partial \sigma_{ij}} = 0,$$

or

$$\frac{\partial L}{\partial \sigma_{ij}} = (1 - \frac{\delta_{ij}}{2}) N \{ \sigma_{ij} - \frac{1}{N} \sum_{k=1}^N X_{ik} X_{jk} + \sum_{s=1}^r \lambda_s \mu_{si} \mu_{sj} \} = 0.$$

These results are summarized in the following:

Theorem 4 The likelihood equations for the parameters of a mixture of multivariate normal distributions with a common covariance matrix are equivalent to equations (5.4), (5.5) and the following:

$$\hat{\sigma}_{ij} = \frac{1}{N} \sum_{k=1}^N x_{ik} x_{jk} - \sum_{s=1}^r \hat{\lambda}_s \hat{\mu}_{si} \hat{\mu}_{sj} \quad (5.7)$$

The equations of Theorems 3 and 4 define a simplified iteration method which is optimal in the limiting case of very widely separated types and which can be accelerated in many other cases by applying equations (3.6) and (3.8) to the mixing proportions and the means. Two computer programs have been written to implement this procedure: the program for Theorem 3 is called NORMIX and the program for Theorem 4 is called NORMAP. Some results of the programs will be presented in section 7.

6. LATENT CLASS ANALYSIS

The general principles of mixture analysis given in Theorems 1 and 2 are applicable to a wide variety of distributions. The first application involved continuous distributions. Next, let us consider the quite different discrete distribution employed in Lazarsfeld's "Latent Class" model for attitude item responses:¹

¹See Anderson (1959) for a discussion of ML estimation and Stanat (1968) for Fourier transform estimation procedures.

$$\alpha_s(x, \mu_s) = \prod_{i=1}^m \mu_{si}^{x_i} (1-\mu_{si})^{1-x_i}, \text{ where } x_i=0 \text{ or } 1. \quad (6.1)$$

$$\frac{\partial \log \alpha_s}{\partial \mu_{si}} = \frac{x_i - \mu_{si}}{\mu_{si}(1-\mu_{si})} \quad (6.2)$$

Substituting this derivative in (2.9), we obtain

$$\frac{\partial \log L}{\partial \mu_{si}} = \frac{-N\lambda_s}{\mu_{si}(1-\mu_{si})} \left\{ \mu_{si} - \frac{1}{N\lambda_s} \sum_{k=1}^N P(s|x_k) x_{ik} \right\} = 0. \quad (6.3)$$

Setting the term in brackets to zero gives us the following remarkable fact.

Theorem 5 The maximum-likelihood estimates for the parameters of a mixture of latent classes are solutions to equations (5.4) and (5.5).

Thus, the estimation equations for this discrete distribution are formally identical to those for a multivariate normal distribution, the only difference being that $\hat{P}(s|x_k)$ is calculated with (6.1) instead of (5.1).

7. EXAMPLES OF COMPUTER ANALYSES

This section presents some results from computer runs with programs NORMAP (normal mixtures with common covariance matrix) and NORMIX (normal mixtures with different covariances matrices).² Both programs are written for flexible input so as to be suitable for general use in a statistical library.

²Both programs are written in FORTRAN 63 for the CDC 1604 and will shortly be available from CO-OP, Control Data Corporation, 3145 Porter Drive, Palo Alto, California 94304.

The programs are currently limited to 10 variables, 1000 individuals, and 20 types. NORMAP is so named because it produces a "map" of the data by generating a printer-plot of the sample points in discriminant-space. In each example to be reported, NORMAP required about one minute to estimate parameters under four hypotheses concerning the number of types, while NORMIX required as much as nine minutes. Several runs were made with different initial estimates. In those cases where multiple solutions were obtained, only the greatest likelihood results were used in the analysis.

7.1 Iris data

The classic Iris data published by Fisher (1936) have been used by Kendall (1965) and Friedman and Rubin (1967) for illustrating their cluster analysis methods. Of course Fisher knew the correct classifications of each of the 150 irises in his sample, but the cluster analysis studies and our mixture analyses attempt to discover and describe the types of irises without using any a priori classification information.

NORMAP and NORMIX were run on the data using hypotheses of one type, two types, three types, and four types. Each hypothesis was tested against the previous one with $\chi^2 = -2 \log(L_{r-1}/L_r)$ and degrees of freedom equal to the difference in the number of parameters in the two hypotheses. The results are given in Table 1.

TABLE 1. FISHER-IRIS SIGNIFICANCE TESTS

No. of Types Null Hypothesis/Alternative	NORMAP		NORMIX	
	χ^2 (d.f.=5)	P	χ^2 (d.f.=15)	P
1/2	166.93	$<10^{-8}$	531.12	$<10^{-8}$
2/3	80.19	$<10^{-8}$	68.54	$<10^{-8}$
3/4	12.04	.034	35.12	.002

The significance tests definitely indicate that the hypothesis of one type should be rejected against the alternative of two types, and that the hypothesis of two types should be rejected against the alternative of three types. It is not quite so clear whether the hypothesis of three types should be rejected against the alternative of four types. Perhaps the obtained values of four types are due to a slight skewness in the component distributions, or perhaps the sample size is so small that the distribution of the log likelihood ratio differs significantly from its asymptotic chi-square distribution. The results seem to show room for improvement in the hypothesis-testing part of the analysis.

Table 2 presents the maximum-likelihood estimates of some of the parameters obtained by NORMAP and NORMIX for three types and compares them with the estimates obtained by Fisher when he used a priori knowledge of the correct classifications. The row labeled "number misclassified" gives the number of sample points whose highest probabilities of membership occurred in types different from their actual species. Both NORMIX and NORMAP gave estimates very close to the Fisher values, with NORMAP slightly better than NORMIX in most cases. The three flowers "misclassified" by NORMAP were identified as numbers 71, 84, and 134. These are precisely the same plants which Friedman and Rubin (1967) misclassified by their $|T|/|W|$ maximization procedure. This is not surprising, because NORMAP could be considered to be a continuous version of the discrete partitioning procedure of Friedman and Rubin. The two methods tend to coincide in the limiting case of widely separated types.

TABLE 2. FISHER IRIS PARAMETERS

PARAMETER	IRIS SETOSA			IRIS VIRGINICA			IRIS VERSICOLOR		
	FISHER	NORMAP	NORMIX	FISHER	NORMAP	NORMIX	FISHER	NORMAP	NORMIX
λ	.3333	.3333	.3333	.3333	.3371	.3670	.3333	.3296	.2990
MEAN 1	5.0060	5.0060	5.0060	6.5880	6.5754	6.5446	5.9360	5.9423	5.9150
MEAN 2	3.4280	3.4280	3.4280	2.9740	2.9807	2.9487	2.7700	2.7608	2.7778
MEAN 3	1.4620	1.4620	1.4620	5.5520	5.5390	5.4797	4.2600	4.2602	4.2017
MEAN 4	.2460	.2460	.2460	2.0260	2.0251	1.9847	1.3260	1.3196	1.2970
NUMBER MISCLASSIFIED		NONE	NONE		1	NONE		2	5

7.2 Artificial clusters

The second example involves three artificially generated clusters consisting of 100, 75, and 50 points in two dimensions. The clusters were deliberately constructed to have multivariate normal distributions with different covariance matrices. Also, the clusters overlap somewhat more than the iris species did. The raw data, scatter plots and some computer printouts are available in a previously published report (Wolfe, 1965).

The results of the hypothesis testing phase of the analysis are given in Table 3. A routine application of the χ^2 likelihood ratio test would indicate the existence of more than three types in the data. However an examination of the parameter estimates for the "fourth" cluster (not shown here) revealed that NORMIX had estimated the fourth cluster's mixing proportion to be .030 - the equivalent of a sample of seven points. The chi-square approximation is inaccurate for this sample size. Evidently, further research is required to develop appropriate significance tests for small samples.³

The parameter estimates for three types are given in Table 4. NORMIX shows a clear superiority in the accuracy of its estimates, as would be expected in a situation where the types have unequal covariance matrices.

³One possibility is a better approximation of the form $\chi^2 = -2 \log Q / (1 + \frac{a}{n})$ where $Q = (L_{r-1} / L_r)$, $n = N \cdot \lambda_{\min}$, and a is a constant to be determined by methods similar to those given by Lawley (1956), or if necessary by Monte Carlo methods. The examples in this paper seem to require a value of $a \sim 10$.

TABLE 5. ARTIFICIAL CLUSTER SIGNIFICANCE TESTS

No. of Types Null Hypothesis/Alternative	NORMAP		NORMIX	
	χ^2_3	P	χ^2_6	P
1/2	25.33	.00001	43.05	.000001
2/3	20.46	.0001	37.20	.00002
5/4	14.04	.003	19.26	.004

TABLE 4. ARTIFICIAL CLUSTER PARAMETERS

PARAMETER	CLUSTER 1			CLUSTER 2			CLUSTER 3		
	ACTUAL	NORMAP	NORMIX	ACTUAL	NORMAP	NORMIX	ACTUAL	NORMAP	NORMIX
λ	.444	.460	.484	.333	.431	.346	.222	.109	.170
MEAN 1	1.41	.92	1.19	.05	.21	.20	-.83	-.49	-1.11
MEAN 2	.85	1.11	.93	-1.33	-1.11	-1.31	1.64	2.50	1.79
S.D. 1	.92	1.30	1.04	1.26	1.30	1.28	.87	1.30	.83
S.D. 2	.97	.69	.91	.49	.69	.50	1.04	.69	1.12
r_{12}	.47	.31	.52	.16	.31	.25	.40	.31	.72
NUMBER MISCLASSIFIED	NONE	22	9	NONE	1	5	NONE	28	12

8. CONCLUDING REMARKS

By reformulating cluster analysis as a problem in estimation for mixed distributions, we hope to have put the subject on a more rigorous foundation. No "similarities" or "distances" need to be assumed a priori. The closest analogy to a "similarity" in our system is the probability of membership of a point in a cluster, but this probability is the result of an iterative solution to the likelihood equations rather than an arbitrarily given function. The partitioning methods of cluster analysis are seen to be approximations to maximum-likelihood solutions, valid when the clusters are widely separated. Mixture analysis is a "smoothed" version of cluster analysis. The difficult combinatorial problems entailed by discrete partitioning methods are avoided by weighting the sample observations with a continuous probability of membership function.

The feasibility of the iterative solution of the maximum-likelihood equations has been demonstrated by the examples in this paper. The iterations involve re-calculating the probabilities of membership of every point in each type. The amount of computation is more extensive than statisticians are accustomed to, but it is of the same order of magnitude as many other cluster analysis systems.

Maximum-likelihood estimation procedures have been used throughout the paper because of the ease with which they can be generalized to multivariate distributions of various forms. The general theory of mixtures has been applied to the multivariate normal and the multivariate Bernoulli distributions. Considerably more work needs to be done in applying the theory to other distributions, in developing significance tests for small

samples, and in finding confidence regions. Further research is desirable in the areas of iteration methods, initial estimation procedures and multiple solutions. Nevertheless, we believe that a good beginning has been made toward the development of logical and practical procedures for the analysis of mixtures.

REFERENCES

1. Anderson, T. W. (1959). "Some scaling models and estimation procedures in the latent class model," pp. 9-38. Probability and Statistics, Edited by Ulf Grenander, John Wiley & Sons, N.Y.
2. Ball, G. H. (1967). "A comparison of two techniques for finding the minimum sum-squared error partition," pp. 7.01-7.47. Conference on Cluster Analysis of Multivariate Data (Defense Documentation Center AD 653 722) Catholic University of America, Washington, D.C.
3. Ball, G. H. and Hall, David J. (1967). PROMENADE, An On-Line Pattern Recognition System, Stanford Research Institute, Technical Report No. RADC-TR-67-310.
4. Blischke, W. R. (1963). "Mixtures of discrete distributions," pp. 351-372, Proceedings of the International Symposium in Classical and Contagious Discrete Distributions, Statistical Publishing Society, Calcutta.
5. Cohen, A. C. (1967). "Estimation in mixtures of two normal distributions," Technometrics, Vol. 9, pp. 15-28.
6. Cooper, Paul W. (1967). "Some topics on nonsupervised adaptive detection for multivariate normal distributions," pp. 123-146, Computer and Information Sciences-II, Edited by Julius T. Tou, Academic Press, N. Y.
7. Fisher, R. A. (1936). "Multiple measurements in taxonomic problems." Annals of Eugenics, Vol. VII, pp. 179-88.
8. Forgy, Edward (1965). "Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications." WNAR meetings, University of California, Riverside, June 22-23, 1965. (see abstract, Biometrics, 21(3) p.768).
9. Friedman, H. P. and Rubin, J. (1967). "On some invariant criteria for grouping data," Journal of American Statistical Association, Vol. 62, pp. 1159-1178.
10. Gibson, W. A. (1959). "Three multivariate models: factor analysis, latent structure analysis, and latent profile analysis," Psychometrika, Vol. 24, pp. 229-252.
11. Hasselblad, Victor (1966). "Estimation of parameters for a mixture of normal distributions." Technometrics, Vol. 8, pp. 431-444.
12. Isaacson, Eugene and Keller, Herbert B. (1966). Analysis of Numerical Methods, John Wiley & Sons, N.Y.

13. Kale, B. K. (1962). "On the solution of likelihood equations by iteration processes. The multiparametric case," Biometrika, Vol. 49, pp. 479-486.
14. Kendall, M. G. (1966). "Discrimination and classification," pp. 165-84. Multivariate Analysis, Edited by P. R. Krishnaiah, Academic Press, N.Y.
15. Lawley, D. N. (1956). "A general method for approximating to the distribution of likelihood ratio criteria." Biometrika, Vol. 43, pp. 295-303.
16. Lazarsfeld, Paul F. (1959). "Latent structure analysis," pp. 476-543, Psychology: A Study of a Science, Vol. 3, Edited by S. Koch, McGraw-Hill, N. Y.
17. Medgyessy, Pal (1961). Decomposition of Superpositions of Distribution Functions, Hungarian Academy of Sciences, Budapest.
18. Pearson, Karl (1894). "Contributions to the mathematical theory of evolution." Philos. Trans. Roy.Soc.London, Vol. 185, pp. 71-110.
19. Rao, C. R. (1952). Advanced Statistical Methods in Biometric Research, John Wiley & Sons, N. Y.
20. Sammon, John W., Jr. (1968). "An adaptive technique for multiple signal detection and identification," pp. 409-439, Pattern Recognition, Edited by Laveen N. Kanal, Thompson Book Co., Washington, D. C.
21. Sokal, R. R. and Sneath, P.H.A. (1963). Principles of Numerical Taxonomy, W. H. Freeman and Co. San Francisco.
22. Stanat, Donald F. (1968). "Unsupervised learning of mixtures of probability functions," pp. 357-389, Pattern Recognition, Edited by Laveen N. Kanal, Thompson Book Co., Washington, D. C.
23. Wolfe, John H. (1965). A Computer Program for the Maximum-likelihood Analysis of Types (Technical Bulletin 65-15), U. S. Naval Personnel Research Activity, San Diego. (Defense Documentation Center AD 620 026).
24. Wolfe, John H. (1967). NORMIX: Computational Methods for Estimating the Parameters of Multivariate Normal Mixtures of Distributions. (Research Memorandum SRM 68-2), U. S. Naval Personnel Research Activity San Diego. (Defense Documentation Center AD 656 588).

UNCLASSIFIED

Security Classification

DOCUMENT CONTROL DATA - R & D

Security Classification of title, body, abstract and indexing annotation is to be entered when the overall report is classified

1. ORIGINATING ACTIVITY (Corporate author): U. S. Naval Personnel Research Activity San Diego, California 92152		2a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED	
		2b. GROUP	
3. REPORT TITLE Pattern Clustering by Multivariate Mixture Analysis			
4. FUNDING NUMBERS (Type of report and the source)			
5. AUTHOR (Last name, middle initial, first name) John H. Wolfe			
6. REPORT DATE		7a. TOTAL NO. OF PAGES 37	7b. NO. OF PAGES 24
8a. CONTRACT OR GRANT NO.		9a. ORIGINATOR'S REPORT NUMBER(S) SRM 69-17	
8b. PROJECT NO. IER160000A03		9b. OTHER REPORT NUMBER: Any other numbers that may be assigned this report.	
10. DISTRIBUTION STATEMENT This document has been approved for public release and sale; its distribution is unlimited.			
11. SUPPLEMENTARY NOTES		12. SECURITY AND MILITARY ACTIVITY Chief of Naval Personnel (Pers-A3) Department of the Navy Washington, D. C. 20370	
13. ABSTRACT Cluster analysis is reformulated as a problem of estimating the parameters of a mixture of multivariate distributions. The maximum-likelihood theory and numerical solution techniques are developed for a fairly general class of distributions. The theory is applied to mixtures of multivariate normals ("NORMIX") and mixtures of multivariate Bernoulli distributions ("Latent Classes"). The feasibility of the procedures is demonstrated by two examples of computer solutions for normal mixture models of the Fisher Iris data and of artificially generated clusters with unequal covariance matrices.			

DD FORM 1473 (1-68)

UNCLASSIFIED

Security Classification

UNCLASSIFIED

Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Normix Normap Mixture Classification Cluster analysis Pattern recognition Statistics Computer program Multivariate analysis						

UNCLASSIFIED

Security Classification